TEXT INDEPENDENT SPEAKER IDENTIFICATION ON NOISY ENVIRONMENTS BY MEANS OF SELF ORGANIZING MAPS

E. Monte, J. Hernando,X.Miró,A. Adolf

Dpt.TSC.Universitat Politécnica de Catalunya Barcelona.Spain
E-mail:enric@tsc.upc.es

ABSTRACT

In this paper we propose a new architecture for speaker recognition. This architecture is independent of the text, robust with the presence of noise, and is based on the Self Organizing Maps (SOM) [1]. We compare the performance of this architecture for different parametrizations, different signal to noise ratios, with another method for speaker identification based on the arithmetic-harmonic sphericity measure on covariance matrices [2],[3].

1. INTRODUCTION

The task of automatic speaker identification consists of labeling an unknown voice as one of a set of known voices. The task can be done within several approaches, either with text dependent recognition o with text independent recognition. The choice of the recognition situation determines the architecture to be used. In the case of text dependent situations a time alignment (DTW) of the utterance with the test can be enough [7], while in the case of text independent situations a probabilistic approach might be more adequate [6]. We decided to limit ourselves to the close set situation, where the problem consists in identifying a speaker from a group of N-known speakers, and to text independent situation.

2. DESCRIPTION OF THE SYSTEM.

2.1. Recognition System based on the SOM Algorithm

The system that we propose, uses the VQ function of the SOM and it's topological property, i.e. the fact that neighboring codewords in the feature space are neighbors in the topological map. In figure 1 we show this property; the contents of a codeword are plotted in the coordinates of that codeword on the SOM. The SOM was tested successfully in the problem of speaker recognition [4], taking only into account the VQ aspect of the SOM. The idea that we propose is to use the SOM as a map labels of codewords. Once the SOM is trained with a database composed by speech material of the speakers that have to be recognized, one can compute the rate of occupancy of each centroid, i.e., the number of times that an input frame is associated to the centroid, and thus make an occupancy histogram. This occupancy histogram is different for each speaker as can seen in figure 2, where we show an example of the histograms for six different speakers. This experiment was done with a SOM of dimensions 10x10, trained with speech material of 100 speakers. The codebook used was the on that is shown in figure 1, and the feature vector consisted of the mel frequency cepstral coefficients.

The computation of the occupancy histogram has an inherent inaccuracy due to the fact that the training and testing material are limited in number. A smoothing of the histograms has revealed to be of use for improving the estimates of these histograms. It can be inferred from figure 1, that if the training material is large enough, the occupancy rate of a codeword will be similar to the occupancy rate of it's neighbors. This smoothing was done using a 2D low pass filtering, which interpolates the value of the occupancy rate of a codeword with the occupancy rate of it's neighboring units. A diagram of the system proposed in this paper is summarized in figure 3. First a training database is used for training the SOM, then for each speaker a occupancy histogram is computed, which is then low pass filtered in order to have a better estimate of the histograms. Once a library of occupancy histograms is trained, the occupancy histograms of the test speakers is computed on the same SOM. Afterwards the distance between the histogram of the test speaker and the histograms in the reference library is computed, and the nearest reference speaker is selected.

The key point of the system is the similarity measure between the occupancy histograms of the test speaker and the reference speaker. This measure has to have a probabilistic interpretation (we approximate the pdf's of the speakers by the histograms). We decided to use the relative entropy, which can be expressed as:

$$RE(S_i, M_j) = \sum_{k=1}^{n} P_k(S_i) \log \frac{P_k(S_i)}{P_k(M_j)}$$

· This paper was supported by the TIC95-1022-C05-03 grant

The relative entropy $RE(S_i, M_j)$ between the test speaker S_i with the model M_j of the reference speaker j, is a function of the occupancy histogram of the test speaker i and the occupancy histogram of the reference model j, where k is a counter that refers to the k-th unit of the map. Thus $P_k(S_i)$ corresponds to the number of times that the k-th unit has been visited in the histogram that corresponds to the test speaker S_i and $P_k(M_j)$ is the equivalent for the reference speaker . This measure has also been proposed in [5], in a system for clustering speakers.

2.2. Recognition System based on the Arithmetic-Harmonic Spheriy Measure

There are a number of techniques that have demonstrated good text independent speaker identification performance in relatively low-noise environments. In this paper, we will compare the system that we propose with a system that uses an arithmetic-harmonic sphericity measure on the covariance matrices of the sequence of the parameter vectors [2],[3], which is easy to implement and computationally efficient. In this system, one reference is used per speaker, which is the covariance matrix of the acoustic parameters of a training utterance.

The arithmetic-harmonic sphericity distance measure between a test covariance matrix **Y** and a reference covariance matrix **X** is defined as:

$$m(\mathbf{X},\mathbf{Y}) = \log\left(\frac{\mathbf{A}}{\mathbf{H}}\right)$$

where **A** and **H** are respectively the arithmetic and harmonic means of the eigenvalues of **Y** relative to **X** (eigenvalues of the product $\mathbf{YX}^{-1}$), that are always positive. This measure is non-negative and equals to zero if **A** = **H**, that is if all eigenvalues are equal (i.e. when **X** and **Y** are proportional). Moreover, m is clearly symmetric. Another important property of this measure comes from the fact that it can be computed very efficiently without an explicit computation of the eigenvalues of **Y** relative to **X**.

That measure is used in association with the 1-nearest neighbor decision rule. The possibility of rejection is not taken into account.

3. SPEECH DATA AND PROCESSING

The database used for testing the system was the TI database. The pre-processing of the signal consisted of a preemphasis of 0.98, the analysis windows had a duration of 30 ms at 10ms rate, the parameter that were computed in different experiments were the LPC with an analysis was of order 24, and the MFCC parameters were of length 24. We used 100 utterances per speaker, and 100 speakers for training a SOM of dimensions 25x25. The same number of utterances were used for computing the occupancy rates. The kernel for the 2D low pass filter for smoothing the occupancy rate had 3x3 coefficients and a normalization of the occupancy histogram was done after the filtering for assuring the stochastic restriction, i.e. that the sum of the elements be one. Noisy speech was simulated by adding zero mean white Gaussian noise to the clean signals. The signals were contaminated by zero mean white Gaussian noise in order that the SNR becomes clean, 30, 20 and 10 dB.

4. RESULTS

The recognition statistics were computed by a test with 55 files per speaker, and the number of speakers was taken to be 100. The results of the test with the SOM were compared with the results obtained by using the arithmetic-harmonic spherity measure(AHSM) on the covariance matrices, and were repeated for several signal noise ratios. The results are presented in table 1, it can be seen that the results obtained by means of the method that uses the SOM are comparable to the results obtained by means of the other method. For high signal to noise ratios the SOM method yields better results, we think that the results can be improved for low signal to noise ratios, when using a more robust parametrization, due to the fact that the method works by computing distances to codewords.

SNR(DB)	clean	30 dB	20 dB	10 dB
LPC/SOM	98,2	95,0	55,0	8,0
MFCC/SOM	100	95,5	53,0	19,5
LPC/AHSM	97,3	85,14	36,5	5,6
MFCC/AHSM	98,6	96,4	68,9	27,6

Table 1: Comparison of the results of the experiment for the two methods

5. SUMMARY

In this paper we have shown that a system based on the statistics of the occupancy rate of the cells of the SOM, can produce recognition results comparable to the results obtained with the arithmetic harmonic spherity measure, for different signal to noise ratios. The results obtained using the SOM method are better for high signal to noise ratio due

to the parametrization, in the future more robust parametrization methods will be used.

6. REFERENCES.

[1] Kohonen,T., "Self Organisation and Associative Memory". Springer-Verlag, 1984.

[2] Bimbot F., Mathan L., "Text-Free Speaker Recognition Using an Arithmetic-Harmonic Sphericity Measure", Proc.EUROSPEECH'93, Belin, Sept.1993.

[3] Hernando J., Nadeu C., Villagrasa C. and Monte E., "Speaker Identification in Noisy Conditions using Linear Prediction of the One-Sided Autocorrelation Sequence",ICSLP 94. Sept 94.Yokohama,Japan

[4] Anderson T., Roy Patterson, " Speaker Recognition with the Auditory Image Model and Self Organization Feature Maps: A Comparison with Traditional techiques", Proc. ESCA wokshop on Automatic Speaker Recognition, Identification, Verfication. Switzerland, April 1994.

[5] Foote J.T., Silverman H.F., "A model distance measure for talker clustering and identification". Proc. ICASSP 94.Adelaide.

[6] Gish, H and Schmidt, M. "Textr-Independent Speaker Identification", IEEE Signal Processing magazine, Oct.94.

[7] Naik, J. "Speaker Verification: A Tutorial", IEEE Communication Magazine, Jan.90.

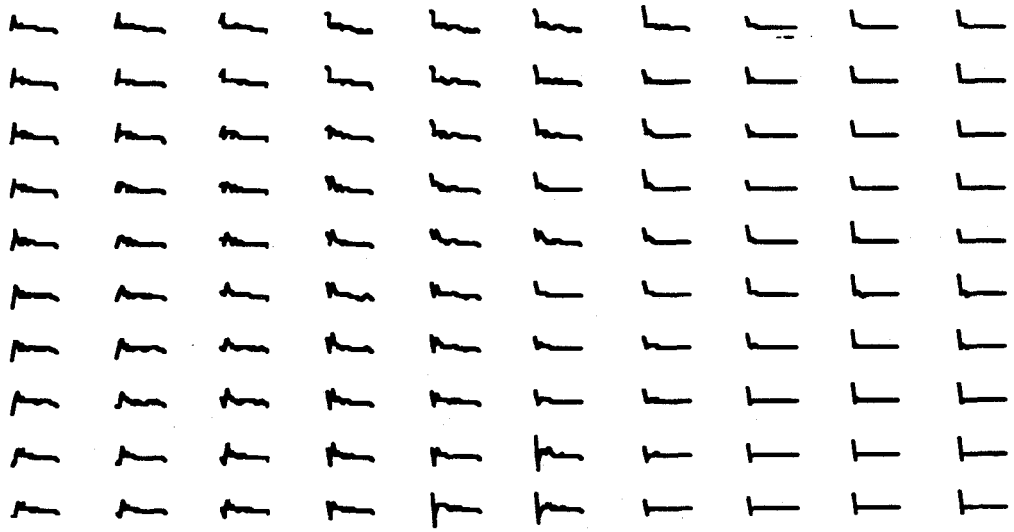

Figure 1:The topological property of the SOM: neighboring units on the SOM are associated with neighboring codewords.

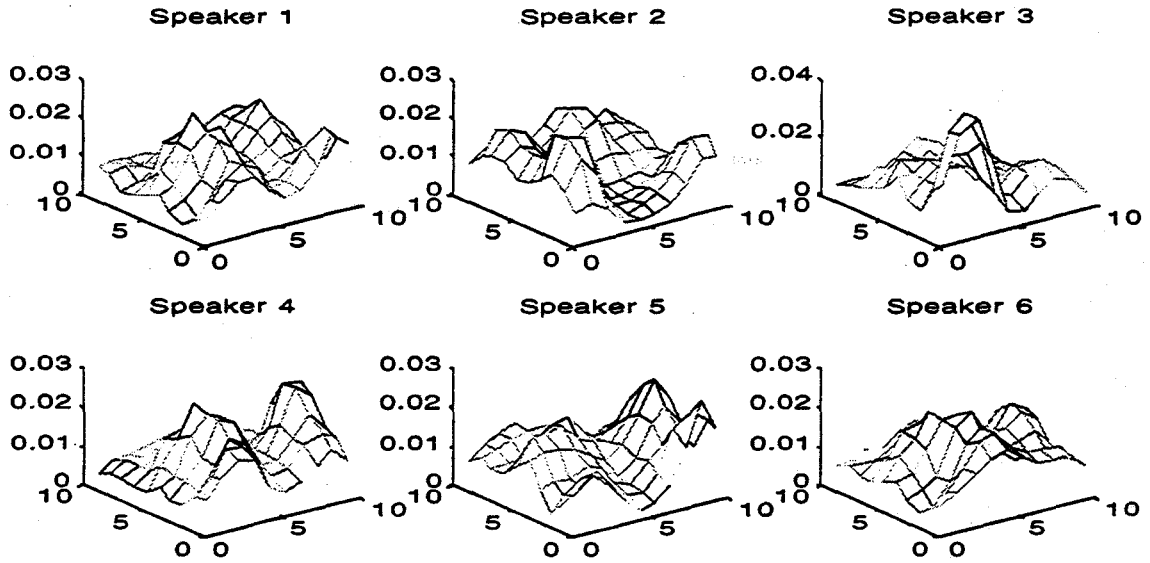

Figure 2. The occupancy histogram of the SOM for six different speakers.

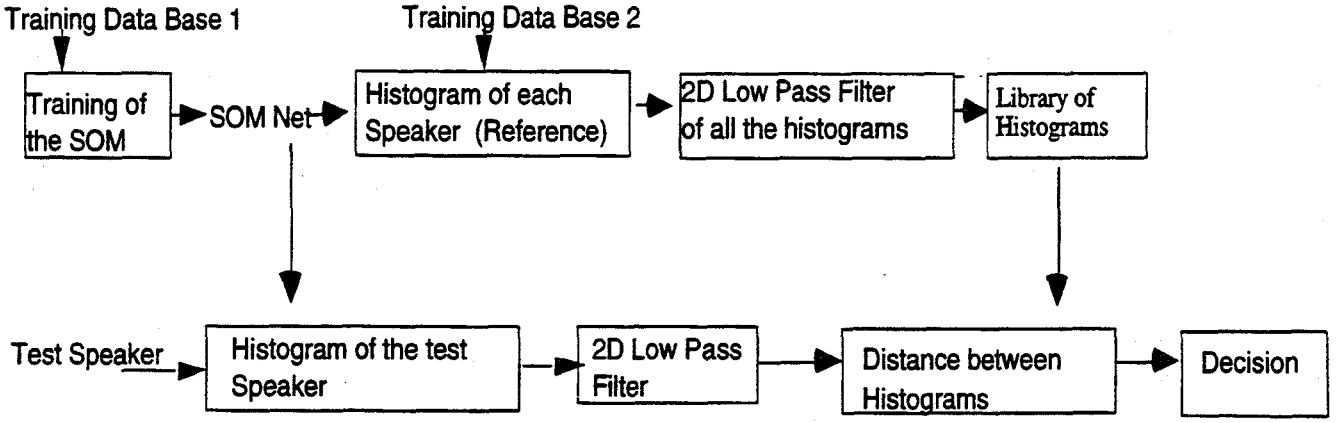

Figure 3. Diagram of the system proposed in this paper.